BAB II

TINJAUAN PUSTAKA DAN DASAR TEORI

2.1 Tinjauan Pustaka

Tinjauan pustaka dalam penelitian ini merupakan referensi penulisan dalam membangun aplikasi. Referensi penelitian ditunjukkan pada Tabel 2.1.

Table 2.1 Tinjauan Pustaka

No	Peneliti	Objek	Metode	Hasil
1	Heru Susanto, Dr. Surya Sumpeno, S.T., M.Sc. , Reza Fuad Rachmadi, S.T., M.T. (2014)	Visualisasi Text Twitter Berbasis Bahasa Indonesia	K-means, *Cascade* K-means, *Self-Organizing Map* (SOM)	Hasil visualisasi data *tweet* terhadap hasil pengklasteran pada 3 variasi algoritma berhasil diimplementasikan pada diagram scatter.
2	Eko Yulian (2018)	Text mining pada tema LGBT dalam arsip *tweet*	K-Means	Dari lima *cluster* yang dibentuk pada proses K-means diperoleh bahwa kecenderungan cuitan pengguna Twitter terkait LGBT secara umum masih berhubungan dangan perspektif religi.
3	Setyo Budi (2017)	Text mining analisis sentimen review film	Algoritma K-Means	Hasil akurasi pengujian dengan menggunakan 300, 700 dan 1000 dokumen dataset mendapatkan nilai *accurary* sebesar 57.83%, 56.71%, dan 50.40%. semakin besar data set yang digunakan maka semakin rendah nilai *accuracy* K-Means.
4	Irfangi (2019)	*Tweet* mengenai transporta-si *online*	*Naïve Bayes Classifier*	Hasil uji akurasi pengujian 109 data, dihasilkan akurasi sebesar 84%.
5	Septian Narsa Putra (2019)	*Tweet* mengenai Divisi Humas Polri	*Naïve Bayes Classifier*	Hasil akurasi pengujian adalah 86%. *Tweet* terbagi menjadi tiga topik: kegiatan polisi, layanan masyarakat dan komentar masyarakat.
6	Usulan (2020)	*Tweet* mengenai kinerja Kabinet Indonesia Maju	K-Means	Sentimen positif, netral dan negatif terhadap *tweet* mengenai kinerja kabinet.

Heru Susanto, Dr. Surya Sumpeno, S.T., M.Sc. , Reza Fuad Rachmadi, S.T., M.T. (2014) melakukan sebuah analisis sentimen dengan menggunakan topik isu Pemilu 2014 dengan data yang digunakan sebanyak 57294 *tweet*. Algoritma pengklasteran yang digunakan adalah K-Means, Cascade K-Means dan *Self-Organizing Map* Kohonen. Hasil yang didapat menunjukkan bahwa *Cascade* K-Means mampu menghasilkan nilai konvergensi kelompok terkecil SSE sebesar 7073 dan Dunn Index 0,67 dengan distribusi sentimen positif berjumlah 26332 *tweet*, negatif berjumlah 7912 *tweet*, dan netral berjumlah 23050 *tweet*.

Eko yulian (2018) melakukan penelitian dengan mengimplementasikan motode K-Means terhadap arsip *tweet* masyarakat kota Bandung dengan tema LGBT, dimana dari lima *cluster* yang dibentuk diperoleh bahwa kecenderungan cuitan pengguna *Twitter* kota bandung terkait LGBT secara umum masih berhubungan dangan perspektif religi. Kemunculan kata agama yang sangat sering menyebabkan asosiasi terhadap kata tersebut cukup besar.

Setyo Budi (2017) melakukan analisis tentang *review* film dengan memanfaatkan data kuisioner yang dikelompokan menjadi 3 bagian yaitu : 1) 300 dokumen review positif dan 300 dokumen review negatif, 2) 700 dokumen review positif dan 700 dokumen review negatif, 3) 1000 dokumen review positif dan 1000 review dokumen negatif. Akurasi yang didapatkan dari penelitian tersebut adalah 57.83% untuk dataset 300 dokumen *review* positif dan 300 dokumen *review* negatif, 56.71% untuk dataset 700 dokumen positif dan 700 dokumen negatif, dan 50.40% untuk dataset 1000 dokumen positif dan 1000 dokumen negatif.

Irfangi (2019) melakukan penelitian dengan mengimplementasikan metode *Naïve Bayes Classifier* untuk melakukan analisis sentimen terhadap trasportasi online di Indonesia pada media Twitter. Hasil uji akurasi pengujian 109 data, dihasilkan nilai akurasi sebesar 84%.

Septian (2019) melakukan penelitian dengan mengimplementasikan metode *Naïve Bayes Classifier* untuk melakukan analisis sentimen pada media Twitter milik Divisi Humas Polri dimana *tweet* yang akan dianalisis diklasifiasikan menjadi tiga topik: kegiatan polisi, layanan masyarakat dan komentar masyarakat. Hasil akurasi pengujian *clustering* pada sistem ini adalah 86%.

2.2 Dasar Teori

2.2.1 Twitter

Twitter merupakan situs jejaring sosial yang membolehkan membaca dan menulis perkembangan terbaru (update), yang dikenal sebagai "*tweet*s". Sistem situs ini berbasis pesan pendek yang ditampilkan pada profil pengguna dan dikirimkan pada pengguna lain yang telah menjalin pertemanan, yang disebut dengan "followers" atau "pengikut" (yudha, 2018).

Layanan seputar Twitter di antaranya :

1. Twitpic, yakni seputar aplikasi untuk mengupload foto dan otomatis memberitahu teman lewat posting di Twitter.
2. Twitterific dan *Tweet*ie adalah applikasi iphone yang memungkinkan pengguna menampilkan dan posting ke Twitter lewat iphone.
3. Twitter Gadget oleh Google, antar muka desktop bagi Twitter.

2.2.2 Twitter API

API atau yang biasa disebut Application Programming Interface adalah suatu program atau aplikasi yang disediakan oleh pihak *developer* tertentu agar kita atau pihak pengembang aplikasi lainnya dapat lebih mudah mengakses aplikasi tersebut tersebut. intinya API ini berfungsi sebagai jembatan antara aplikasi satu dengan aplikasi yang lain (Bramanda, 2014).

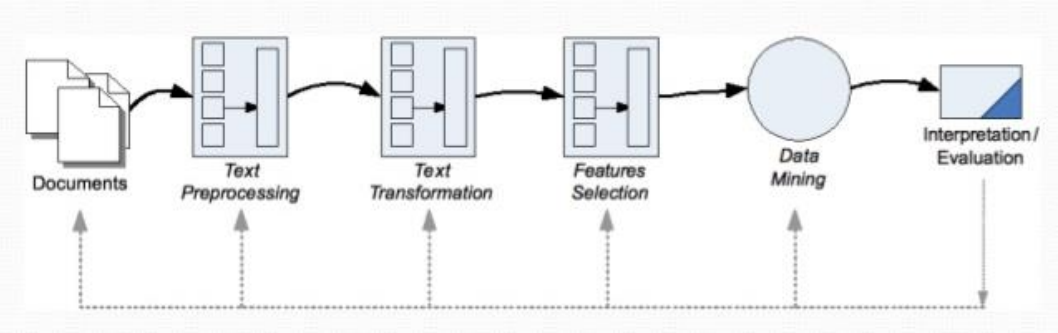
Twitter API yaitu sebuah aplikasi yang diciptakan oleh pihak Twitter agar mempermudah pihak *developer* lain untuk mengakses Informasi web Twitter tersebut (Bramanda, 2014). Untuk mengakses API tersebut dibutuhkan kunci (*Consumer Key dan Consumer Secret*).

2.2.3 Text Mining

Text mining dapat didefinisikan secara luas sebagai proses pengetahuan intensif dimana pengguna berinteraksi dengan koleksi dokumen dari waktu ke waktu dengan menggunakan seperangkat alat analisis. Text mining berusaha mengekstrak informasi yang berguna dari sumber data melalui identifikasi dan eksplorasi pola yang menarik.

Text mining cenderung mengarah pada bidang penelitian data mining. Oleh karena itu, tidak mengherankan bahwa text mining dan data mining berada pada tingkat arsitektur yang sama. Penambangan teks dapat dianggap sebagai proses dua tahap yang diawali dengan penerapan struktur terhadap sumber data teks dan dilanjutkan dengan ekstraksi informasi dan pengetahuan yang relevan dari data teks terstruktur ini dengan menggunakan teknik dan alat yang sama dengan penambangan data (Fahlahah, 2015).

Gambar 2.1 menunjukkan empat tahapan proses dalam text mining yang terdiri dari pemrosesan awal terhadap teks (*text preprocessing*), transformasi teks (*text transformation*), pemilihan fitur (*feature selection*), dan penemuan pola *(pattern discovery*) (Eko, 2011).



Gambar 2.1 Tahapan Text Mining

Pertama, *Text Preprocessing*. Tahap ini melakukan analisis semantik (kebenaran arti) dan sintaktik (kebenaran susunan) terhadap teks. Tujuan dari pemrosesan awal adalah untuk mempersiapkan teks menjadi data yang akan mengalami pengolahan lebih lanjut. Operasi yang dapat dilakukan pada tahap ini meliputi part-of-speech (PoS) tagging, menghasilkan parse tree untuk tiap-tiap kalimat, dan pembersihan teks.

Kedua, *Text Transformation*. Transformasi teks atau pembentukan atribut mengacu pada proses untuk mendapatkan representasi dokumen yang diharapkan. Pendekatan representasi dokumen yang lazim digunakan oleh model “bag of words” dan model ruang vector (vector space model). Transformasi teks sekaligus juga melakukan pengubahan kata-kata ke bentuk dasarnya dan pengurangan dimensi kata di dalam dokumen. Tindakan ini diwujudkan dengan menerapkan stemming dan menghapus stop words.

Ketiga, *Feature Selection*. Pemilihan fitur (kata) merupakan tahap lanjut dari pengurangan dimensi pada proses transformasi teks. Walaupun tahap sebelumnya sudah melakukan penghapusan kata-kata yang tidak deskriptif (stopwords), namun tidak semua kata-kata di dalam dokumen memiliki arti penting. Oleh karena itu, untuk mengurangi dimensi, pemilihan hanya dilakukan terhadap kata-kata yang relevan yang benar-benar merepresentasikan isi dari suatu dokumen.

Keempat, *Pattern Discovery*. *Pattern discovery* merupakan tahap penting untuk menemukan pola atau pengetahuan (*knowledge*) dari keseluruhan teks. Tindakan yang lazim dilakukan pada tahap ini adalah operasi text mining, dan biasanya menggunakan teknik-teknik data mining. Dalam penemuan pola ini, proses text mining dikombinasikan dengan proses-proses data mining. Masukan awal dari proses text mining adalah suatu data teks dan menghasilkan keluaran berupa pola sebagai hasil interpretasi atau evaluasi. Apabila hasil keluaran dari penemuan pola belum sesuai untuk aplikasi, dilanjutkan evaluasi dengan melakukan iterasi ke satu atau beberapa tahap sebelumnya. Sebaliknya, hasil interpretasi merupakan tahap akhir dari proses text mining dan akan disajikan ke pengguna dalam bentuk visual (Eko, 2011).

2.2.4 Analisis Sentimen

Analisa sentimen atau biasa disebut opinion mining merupakan salah satu cabang penelitian Text Mining. Opinion mining adalah riset komputasional dari opini, sentimen dan emosi yang diekspresikan secara tekstual. Jika diberikan suatu set dokumen teks yang berisi opini mengenai suatu objek, maka opinion mining bertujuan untuk mengekstrak atribut dan komponen dari objek yang telah di berikan

komentar pada setiap dokumen dan untuk menentukan apakah komentar tersebut bermakna negatif atau positif (Falahah, 2015).

2.2.5 *Clustering*

Teknik *clustering* termasuk ke dalam teknik unsupervised learning dimana kita tidak perlu melatih metode tersebut atau dengan kata lain, tidak ada fase pembelajaran *(learning)*. Santosa (2007) menjelaskan bahwa teknik unsupervised learning adalah metode-metode yang tidak membutuhkan label ataupun keluaran dari setiap data yang diinvestigasi. Tujuan utama dari *clustering* adalah pengelompokan objek-objek yang mirip kedalam satu klaster dan berusaha membuat jarak antar klaster sejauh mungkin. Tingkat kemiripan objek-objek dalam satu klaster dapat dilihat dengan membandingkan jarak objek ke centroid satu dengan centroid lainnya. Terdapat beberapa metode yang sering digunakan untuk pencarian jarak, diantaranya Manhattan dan Euclidean. Euclidean sering digunakan karena penghitungan jarak dalam distance space merupakan jarak terpendek yang bisa didapatkan antara dua titik yang diperhitungkan, sedangkan Manhattan sering digunakan karena kemampuannya dalam mendeteksi keadaan khusus seperti keberadaaan outliers dengan lebih baik.

2.2.6 *Preprocessing*

Tahap *preprocessing* atau praproses data merupakan proses untuk mempersiapkan data mentah sebelum dilakukan proses lain. Pada umumnya, praproses data dilakukan dengan cara mengeliminasi data yang tidak sesuai atau mengubah data menjadi bentuk yang lebih mudah diproses oleh sistem. Praproes sangat penting dalam melakukan analisis sentimen, terutama untuk media sosial

yang sebagian besar berisi kata - kata atau kalimat yang tidak formal dan tidak terstruktur serta memiliki noise yang besar (A Clark, 2003).

Pada Tahap *preprocessing* memiliki beberapa tahap proses sebagai berikut:

- *Lower Case conversion* (konversi hurup kecil), langkah pra-pemrosesan pertama yang dilakukan adalah mengubah semua teks menjadi huruf kecil. Ini menghindari memiliki banyak salinan dari kata yang sama pada teks.
- *Removing Punctuations* (menghapus tanda baca), bertujuan untuk menghapus semua karakter misalnya simbol, tanda baca dan lain-lain.
- Stop Words Removal - Kata-kata yang umum muncul harus dihapus dari data teks.
- *Rare Words Removal* diproses pada sebuah kalimat jika mengandung kata-kata yang sering keluar dan di anggap tidak penting seperti waktu, penghubung, dan lain sebagainya sehingga perlu dilakukan penghapusan.
- *Tokenization* - Tokenisasi mengacu pada membagi teks menjadi urutan kata atau kalimat.
- *Lemmatization* / *Stemming*, yaitu mengkonversi kata menjadi kata dasar, bukan hanya menghapusnya saja. Itu menggunakan kosa kata dan melakukan analisis morfologis untuk mendapatkan kata dasar.

2.2.7 Pembobotan TF-IDF

Term Frequency - Inverse Document Frequency atau TF-IDF adalah suatu metode algoritma yang berguna untuk menghitung bobot setiap kata yang umum digunakan. Metode ini juga terkenal efisien, mudah dan memiliki hasil yang akurat. Metode ini akan menghitung nilai Term Frequency (TF) dan Inverse Document Frequency (IDF) pada setiap token (kata) di setiap dokumen dalam korpus. Secara

sederhana, metode TF-IDF digunakan untuk mengetahui berapa sering suatu kata muncul di dalam dokumen. Pada Term Frequency (TF), terdapat beberapa jenis formula yang dapat digunakan (Delta, 2019) :

1. TF biner (binary TF), hanya memperhatikan apakah suatu kata atau term ada atau tidak dalam dokumen, jika ada diberi nilai satu (1), jika tidak diberi nilai nol (0).
2. TF murni (raw TF), nilai TF diberikan berdasarkan jumlah kemunculan suatu term di dokumen. Contohnya, jika muncul lima (5) kali maka kata tersebut akan bernilai lima (5).
3. TF normalisasi, menggunakan perbandingan antara frekuensi sebuah term dengan nilai maksimum dari keseluruhan atau kumpulan frekuensi term yang ada pada suatu dokumen.
4. TF logaritmik, hal ini untuk menghindari dominansi dokumen yang mengandung sedikit term dalam query, namun mempunyai frekuensi yang tinggi.

$$TF = \begin{cases} 1 + log_{10}(f_{t,d}), & f_{t,d} > 0 \\ 0, & f_{t,d} = 0 \end{cases} \quad \text{...(i)}$$

IDF (Inverse Document Frequency). Metode IDF merupakan sebuah perhitungan dari bagaimana term didistribusikan secara luas pada koleksi dokumen yang bersangkutan. Berbeda dengan TF yang semakin sering frekuensi kata muncul maka nilai semakin besar, dalam IDF, semakin sedikit frekuensi kata muncul dalam dokumen, maka makin besar nilainya (Delta, 2019).

$$IDF_j = \log(D/df_j) \quad \text{...(ii)}$$

Dimana *D* adalah jumlah semua dokumen dalam koleksi sedangkan *dfj* adalah jumlah dokumen yang mengandung *term (tj).*

Jenis formula TF yang biasa digunakan untuk perhitungan adalah TF murni (raw TF). Dengan demikian rumus umum untuk Term Weighting TF-IDF adalah penggabungan dari formula perhitungan raw TF dengan formula IDF dengan cara mengalikan nilai TF dengan nilai IDF :

$$w_{ij} = tf_{ij} \times idf_j$$
$$w_{ij} = tf_{ij} \times \log\ (D/df_j) \quad \ldots(\text{iii})$$

Dimana *Wij* adalah bobot *term* (*tj*) terhadap dokumen (*di*). Sedangkan *tfij* adalah jumlah kemunculan *term (tj)* dalam *dokumen (di)*. *D* adalah adalah jumlah semua dokumen yang ada dalam database dan *dfj* adalah jumlah dokumen yang mengandung *term (tj)* (minimal ada satu kata yaitu *term (tj)*).

Berapapun besarnya nilai *tfij*, apabila *D = dfj*, maka akan didapatkan hasil 0 (nol), dikarenakan hasil dari *log1*, untuk perhitungan IDF. Untuk itu dapat ditambahkan nilai 1 pada sisi IDF, sehingga perhitungan bobotnya menjadi sebagai berikut :

$$w_{ij} = tf_{ij} \times \log(D/df_j) + 1 \quad \ldots(\text{iv})$$

2.2.8 K-Means

Algoritma K-Means merupakan salah satu jenis *clustering*. Seperti algoritma unsupervised lainnya, algoritma k-means menerima masukan berupa data tanpa label kelas. Algoritma k-means mengelompokan data-data kedalam k kelompok, dimana kelompok tersebut dibentuk dengan meminimalkan jumlah dari Euclidean distance diantara data dengan titik pusat (*centroid*) yang mempresentasikan *cluster*

atau kelompok tersebut (Teguh, 2018). Berikut adalah tahapan-tahapan dalam algoritma K-Means *clustering* :

1. Menentukan jumlah *cluster* K dari dataset yang akan dibagi.
2. Menentukan data k yang menjadi titik pusat atau *centroid* awal lokasi cluster.
3. Mengelompokkan data ke dalam K *cluster* sesuai dengan titik *centroid* terdekat yang telah ditentukan sebelumnya.
4. Memperbaharui nilai titik centroid dan mengulangi langkah 3 sampai nilai dari titik centroid tersebut tidak berubah.

Sesuai dengan langkah di atas, maka pengelompokkan data dilakukan dengan cara menghitung jarak terdekat data ke centroid sehingga akhirnya membentuk sebuah cluster. Untuk menghitung centroid fitur ke-i digunakan formula :

$$C_i = \frac{1}{M} \sum_{j=1}^{M} x_j \ldots \text{(v)}$$

Berikut adalah beberapa formula yang digunakan untuk menghitung jarak, pengukuran jarak pada ruang jarak (distance space) Euclidean menggunakan formula :

$$D(x_2, x_1) = \|x_2 - x_1\|_2 = \sqrt{\sum_{j=1}^{p} |x_{2j} - x_{1j}|^2} \ldots \text{(vi)}$$

Pengukuran jarak pada ruang jarak Manhattan menggunakan formula :

$$D(x_2, x_1) = \|x_2 - x_1\|_1 = \sum_{j=1}^{p} |x_{2j} - x_{1j}| \ldots \text{(vii)}$$

Pengukuran jarak pada ruang jarak Minkowsky menggunakan formula :

$$D(x_2, x_1) = \|x_2 - x_1\|_\lambda = \sqrt[\lambda]{\sum_{j=1}^{p} |x_{2j} - x_{1j}|^\lambda} \ldots \text{(viii)}$$

Di mana :

D = jarak Antara data x_2 dan x_1

x_2 , x_1 adalah dua buah data yang akan dihitung jaraknya.

λ = parameter jarak Minkowsky.

Selanjutnya pembaharuan suatu titik *centroid* dapat dilakukan dengan formula :

$$\mu_k = \frac{1}{N_k} \sum_{q=1}^{N_k} x_q \quad \ldots(\text{ix})$$

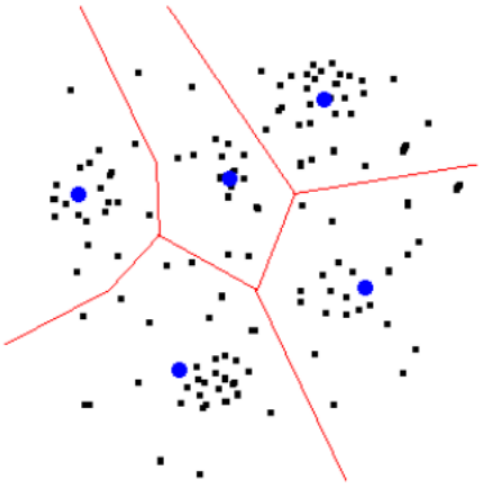
Di mana :

μ_k = titik *centroid* dari *cluster* ke-K.

N_k = banyaknya data pada *cluster* ke-K.

x_q = data ke-q pada *cluster* ke-K

Gambar 2.2 menunjukkan ilustrasi k-means. Titik hitam menyatakan data. Garis merah menyatakan partisi/pemisah. Titik biru merepresentasikan titik pusat (centroid) yang mendefinisikan suatu partisi



Gambar 2.2 Ilustrasi K-Means